Exploring Synthetic Data Generation on Extremely Low-Resource Clinical Temporal Relation Extraction

Exploración de la generación de datos sintéticos en la extracción de relaciones temporales clínicas con recursos extremadamente limitados

Edgar Andrés Santamaria[1], **Oier López de Lacalle**[2]
Aitziber Atutxa Salazar[1]
[1]Bilbao School of Engineering, [2]San Sebastian Faculty of Computer Science,
HiTZ Basque Center for Language Technology
University of the Basque Country (EHU), Spain
{edgar.andres, oier.lopezdelacalle, aitziber.atucha}@ehu.eus

Abstract: Temporal relation extraction (TRE) in clinical texts is essential but challenging due to limited annotated data and the domain-specific nature of medical language. This paper explores diversity/fidelity trade-off in data augmentation using large language models to generate synthetic training data under minimal supervision. We compare two approaches, Triplet-based Generation (TBG) and Paraphrase-based Generation (PBG), and evaluate their effectiveness across discriminative models. Our findings show that minimal manual annotation, even as little as one document, can significantly improve domain relevance and model performance. These results highlight the importance of supervision and data augmentation strategy selection when applying LLMs to domain-specific tasks like clinical TRE.
Keywords: Clinical Temporal Relation Extraction, extremely scarce data regimes, Data Augmentation Strategies

Resumen: La extracción de relaciones temporales (TRE) en textos clínicos es esencial, pero supone un reto debido a la escasez de datos anotados y a la naturaleza específica del lenguaje médico. Este artículo explora el equilibrio entre diversidad y fidelidad en escenarios de aumento de datos utilizando Modelos de lenguaje (LLM) para generar datos de entrenamiento sintéticos con una supervisión mínima. Comparamos dos enfoques, la generación basada en tripletas (TBG) y la generación basada en paráfrasis (PBG), y evaluamos su eficacia en modelos discriminativos. Nuestros hallazgos muestran que una anotación manual mínima, incluso de un solo documento, puede mejorar significativamente la relevancia del dominio y el rendimiento del modelo. Estos resultados ponen de relieve la importancia de la supervisión y la selección de estrategias de aumento de datos a la hora de aplicar los LLM a tareas específicas de un dominio, como TRE clínica.
Palabras clave: Extracción de relaciones temporales clínicas, regímenes de datos extremadamente escasos, estrategias de aumento de datos.

1 Introduction

Temporal information is crucial for healthcare providers to make the right medical decisions; health conditions change over time, worsening, improving, or remaining stable, symptoms tend to appear in a specific sequence, monitoring diagnostic tests and procedures over time enables clinicians to assess treatment effectiveness and identify or prevent adverse outcomes related to side effects. The critical role of temporal information in medical decision-making has been widely recognized. Prior research demonstrates that temporal representations derived from sequences of electronic health record (EHR) events outperform non-temporal data in predictive and classification tasks, such as phenotype prediction (Estiri et al., 2020). In chronic disease risk prediction, capturing temporal relationships is essential be-

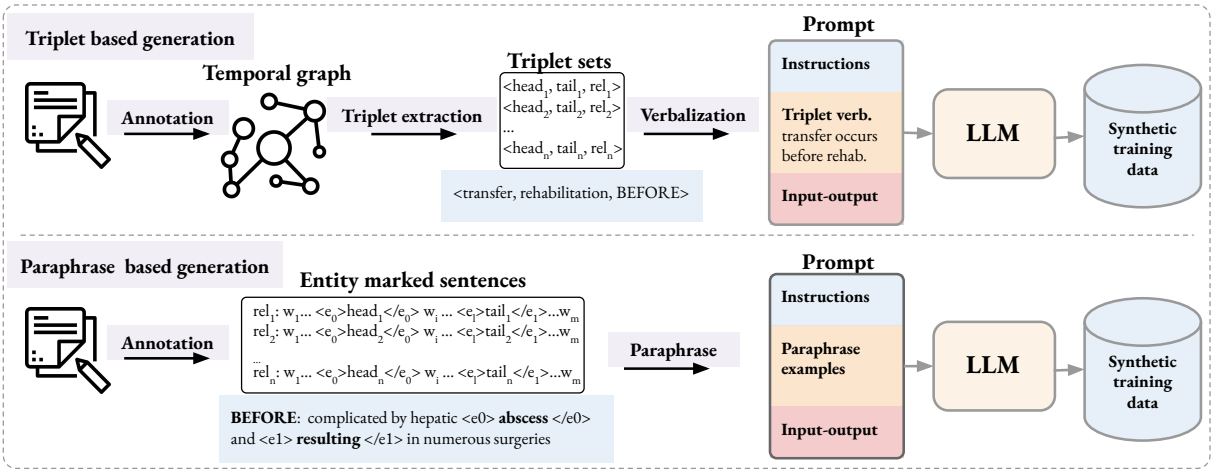


Figure 1: Synthetic data generation pipelines. The top pipeline illustrates the **triplet-based generation** (TBG) methodology. Bottom pipeline is the **paraphrase-based generation** (PBG).

cause certain events are relevant only within specific temporal contexts (Sheikhalishahi et al., 2019). Studies on conditions such as sepsis (Zhu et al., 2021) and lupus erythematosus further show that temporal modeling enhances performance, interpretability, and consistency compared to non-temporal approaches (Wang et al., 2020). For example, Ross et al. (Ross et al., 2010) reported that 40% of clinical trial inclusion criteria exhibit a dependency on temporal data. Clinical Temporal Relation Extraction (CTRE) is a specific case of Relation Extraction (RE), a well-known NLP task, which involves identifying and extracting relationships between entities in text. In this case, the entities are clinical events and time expressions constrained by temporal logic within a specific clinical history progression.

Unlike general RE, CTRE poses two major challenges. First, every relevant entity, whether an event or a time expression, is by definition temporally related or ordered in relation to every other within an implicit timeline. In practice, however, most of these pairwise relationships, although implicitly connected through temporal logic, are neither explicitly annotated in the data (Ning, Feng, and Roth, 2017) nor overtly expressed in the text. Many of these relations can be inferred through reasoning, such as transitivity (e.g., if A occurs before B and B before C, then A occurs before C), which further reduces the need for explicit mention or annotation. Second and most importantly, clinical data introduces practical constraints. It is sensitive, tightly regulated, and requires domain expertise for annotation. As a result, data availability is limited and cross-domain transfer is challenging due to inconsistencies in annotation schemas. Adapting to a new corpus often requires developing a custom schema and manually annotating it from scratch. All these leading to data scarce scenarios (Alfattni, Peek, and Nenadic, 2020; Olex and McInnes, 2022). To address annotated corpus limitations, Data Augmentation (DA) is a commonly employed strategy, as it has been shown to improve performance in data-scarce settings (Wei and Zou, 2019; Xia et al., 2019; Gao et al., 2023; Ghosh et al., 2023; Zhong et al., 2025). However, DA in clinical data-scarce settings presents several challenges. On one hand, key issues include maintaining distributional fidelity (to avoid unrealistic or misleading samples), balancing diversity with domain relevance, ensuring semantic consistency, and avoiding the amplification of existing biases. On the other hand, maintaining temporal graph coherence requires that temporally related pairwise entities fit into a unified, connected timeline. State-of-the-art approaches, such as SynthIE (Josifoski et al., 2023), are not viable for this task due to the lack of a comprehensive temporal knowledge base that covers both clinical and general events.

We leverage the capabilities of contemporary LLMs to generate synthetic data for temporal relation extraction in the clinical domain, with two primary objectives: (1) to address the scarcity of manually anno-

tated corpora, and (2) to mitigate issues of temporal coherence. To this end, we propose two synthetic data generation approaches. The first relies on partially informed settings, where small sets of real triplets (*entity1*, *entity2*, and *relation*) are used as seeds without additional contextual information for augmentation. The second considers fully specified examples that incorporate the surrounding context (*entity1*, *entity2*, and *context*_containing the relation_). An overview of the data generation pipelines is presented in Figure 1. We further evaluate the impact of different augmentation strategies on the ability of CTRE models to transfer knowledge across the THYME and E3C temporally annotated medical datasets, to identify the most effective and robust approaches. As a result, we find that:

- Minimal supervision is crucial for effectively controlling both the diversity and domain relevance of generated data. In contrast, zero-shot generation fails to ensure domain adherence, often producing content that drifts away from the target domain.
- Annotating a single document can significantly improve the domain relevance and quality of the resulting synthetic data.
- Generating text conditioned only on entity pairs and relation labels proves as effective as including full contextual information when manual annotation is extremely limited.
- Synthetic data provides robustness across evaluation datasets. Models informed by one dataset (e.g., E3C) remain competitive when evaluating in the other dataset (e.g., THYME).

2 Related Work

Since the irruption of the transformer architectures (Vaswani et al., 2017), approaches to solve RE predominantly rely on pre-trained LLMs like BERT (Devlin et al., 2019) and special token embeddings to facilitate classification (Classical-BERT-based RE). Works on these lines specifically in Temporal RE include (Ning, Subramanian, and Roth, 2019; Lin et al., 2019; Zhou et al., 2021; Lin et al., 2021; Lin et al., 2023; Knez and Žitnik, 2024). (Lin et al., 2019) that introduces the use of special tokens to mark events (Baldini Soares et al., 2019) and employs a BERT network, using the [CLS] token embedding to classify temporal relations. (Zhou et al., 2021) improves performance through soft logic regularization, predicting relation probabilities while ensuring consistency with rule-based constraints.

CTRE follows a similar path. (Dligach et al., 2017) utilized recurring neural networks (RNNs) such as LSTM to automatically learn features for temporal relation classification. (Lin et al., 2021; Cheng and Weiss, 2023) enhances extraction in the medical domain by utilizing entity-specific pre-training of BERT. Finally, (Knez and Žitnik, 2024), following the work done by (Lin et al., 2023) for medical entity factual relations, explores a bimodal architecture integrating information from text documents and knowledge graphs (Chaturvedi et al., 2025), combining them to make a unified prediction.

Recent research has investigated the use of LLMs for general RE (Jimenez Gutierrez et al., 2022; Delmas, Wysocka, and Freitas, 2024; Labrak, Rouvier, and Dufour, 2024), temporal RE for identifying Temporal Links (TLINK) (Roccabruna, Rizzoli, and Riccardi, 2024), and specialized CTRE (El Khettari, Quiniou, and Chaffron, 2025). Although these studies evaluate generative models using approaches such as in-context learning and LoRA, they consistently find that discriminative models achieve higher extraction accuracy. (Roccabruna, Rizzoli, and Riccardi, 2024) attributes this performance gap to differences in pretraining objectives, suggesting that generative models tend to overemphasize the final token, thereby making limited use of the full input context.

Overall, these findings indicate that, given the substantial computational cost of fine-tuning, generative models may not be the most effective choice compared to discriminative alternatives. However, discriminative methods are themselves constrained by their reliance on large annotated training data, which is an important limitation in domains such as healthcare, where labeled data is scarce.

While DA has been widely adopted in computer vision, its application to RE/TRE is a growing area. Early methods for synthetic data generation focused mainly on rule-based or statistical sampling techniques.

For tabular data, (Rubin and McHugh, 1987)´s work on multiple imputation is a foundational concept; other methods include synonym replacement, random word insertion, deletion, or swapping (Wei and Zou, 2019). More advanced approaches leverage LLM models to generate synthetic examples. For instance, (Josifoski et al., 2023) shows for RE that synthetic data generated using LLMs can resolve data scarcity in low-resource scenarios and mitigate domain disparity compared to distant supervision. SynthIE (Josifoski et al., 2023) proposes methods to synthesize diverse and high-quality training instances for information extraction tasks. However, their approach relies heavily on the availability of comprehensive knowledge bases like Wikidata, which provide entity-relation triplets in the general domain. In our case, no such knowledge base exists that covers temporal relations, especially those involving medical entities. Additionally, their evaluation is affected by dataset bias, likely leading to an overestimation of model performance. This is because both the training and test sets were generated through DA using the same LLM, resulting in the test set closely mirroring the structure and distribution of the training set.

3 Data Augmentation for CTRE

Synthetic data generation aims to produce artificial datasets that replicate the statistical characteristics of real-world data while minimizing inherent biases. In this case, the task involves the participating entities (*head*, *tail*), the temporal relation connecting them, and the contextual information in which this relation is expressed. We use the *head*, *tail* notation because temporal relations are directional, and each of the proposed strategies relies on three components that include the directionality of the relation type.

To this end, we investigated two main distinct strategies for generating new examples by means of LLM prompting as proposed in Figure 1. These two strategies differ in the extent to which LLM prompts are informed by the key components of the original sentences. The first strategy described in Section 3.1, Triplet-based Generation (TBG), is the least informed. The prompt relies solely on entities and the relation between them. This strategy aims to be more creative introducing more diversity, although somehow controlled by providing the real entities and relations. The second strategy explained in section 3.2, Paraphrase-based Generation (PBG) relies on entities and the surrounding context, aims to be more faithful to the original domain distribution and therefore less creative and more faithful to the original.

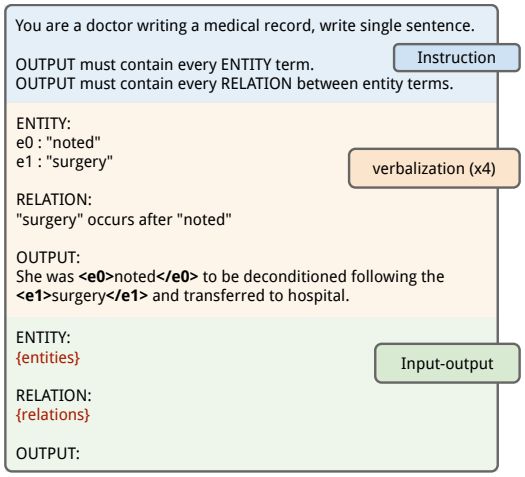


Figure 2: Prompt schema for the triplet-based generation (TBG).

Modeling non-existent relations (i.e., `OUTREL` relation type) is inherently challenging. We chose to model them as inverse relations of the original ones, creating the `OUTREL`´s by reversing the direction of the *head* and *tail* entities while preserving the original sentence context. It is worth noting that the `OVERLAP` relation does not have a meaningful inverse, since if A overlaps B, the reverse (B overlaps A) also holds. We retained only those output instances that preserved the original entities, discarding any cases where the LLM failed to maintain original entities.

3.1 Triplet-based Generation

This section presents the TBG method for generating synthetic clinical text by leveraging real triplets of information (*head*, *tail*, *temporal relation*) derived from manually annotated temporal graphs. In contrast to approaches that rely on broad contextual cues or paraphrasing existing sentences, the triplet-based generation strategy focuses on directly incorporating relationships within predefined verbalizations, as described in (Santamaria et al., 2025). These verbalizations shown in Table 8 are then used to prompt LLMs to generate newly constructed contexts, Figure 1 (top) depicts the data gen-

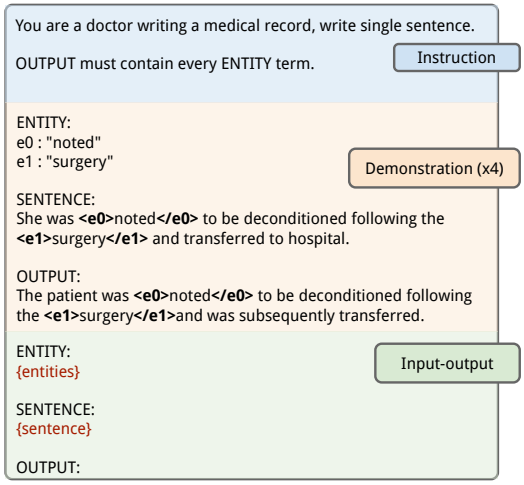


Figure 3: Prompt schema for the Paraphrase-based generation (PBG).

eration workflow.

For each triplet in the gold graph, we generate synthetic examples using a few-shot prompt that verbalizes the temporal relation. As shown in Figure 2, the prompt instructs the LLM to produce sentences containing the original entities and a natural language description of their relationship (e.g., ´surgery occurs after noted' for the `BEFORE` label). We use high-temperature sampling to encourage linguistic diversity, subsequently filtering out any outputs that fail to correctly mark the required entities. `OUTREL` instances are generated by inverting the entity marks order of `CONTAINS` and `BEFORE` relations while preserving the original synthetic context.

The TBG is particularly valuable for augmenting an existing dataset by creating a new one, ensuring faithfulness towards specific entities and relational instances, and simultaneously generating sufficient contextual diversity. The core principle involves providing an LLM with a clear directive, as shown in Figure 2, relying on LLM semantics knowledge to create suitable contexts.

3.2 Paraphrase-based Generation

This section introduces PBG strategy designed to create synthetic clinical text based on a triplet (*head*, *tail*, *context*), and focuses on rephrasing the context to maintain the original semantic relation between two entities. Contrary to TBG method which does not use the original context the core idea is to provide the LLM with the original sentence (*context*), explicitly marking entities to produce a new context that conveys the same relationship between the entities, but different wording and grammatical structure.

Corpus	Scen.	NEG	CON	BEF	OVE	POS
E3C	1-doc	202	23.4	5.8	11.8	41
	2-doc	433.2	56.2	12	17.4	85.6
	10-doc	1544.8	174.8	44.4	75	294.2
	FT	11024	1027	262	617	1906
THYME	1-doc	213.6	49.8	16	21	86.8
	2-doc	251.6	66.8	17.8	13	97.6
	10-doc	2154.8	519	106	113.2	738.2
	FT	56617	8659	1829	2389	12877

Table 1: Average number of gold annotations per positive labels (`CONTAINS`, `BEFORE`, and `OVERLAP`), and negative labels (`OUTREL`) across the different data-scarse scenarios (1-doc, 2-doc, 10-doc) and full-traning setting (FT). POS stands for the total number of positive label annotations.

The designed prompt is shown in Figure 3, and follows a similar structure used for the tripled-based generation. Apart from the same instructions, few-shot examples contain marked entities, marked original sentence, and valid outputs. The `OUTREL` examples are generated in the same way as for TBG generation. For each paraphrase candidate, we construct a prompt that includes the original sentence to generate synthetic examples corresponding to the original triplet. The generation process is configured with a temperature of 0 to preserve the original meaning. Multiple candidates are produced, and those that fail to retain the marked entities within the context are discarded. This process is repeated until the desired number of unique examples is obtained.

4 Experimental Settings

4.1 Datasets

As mentioned in the introduction, CTRE corpora are scarce. For this work, we used THYME and E3C corpora distributed as shown in Table 1. THYME (Styler IV et al., 2014) is built upon de-identified real colon cancer medical records sourced from discharge summaries, ensuring its relevance and authenticity for clinical research. E3C corpus (Magnini et al., 2021), composed by general clinical statements presenting the reason for a clinical visit, the description of physical exams, and the assessment of the patient's situation.

Both THYME and E3C contain annota-

tions of various elements within the health records:

Events: These include significant medical occurrences such as "MRI", "surgery", "biopsy", "chemotherapy session" and "doctor´s consultation". These events form the crucial anchors of the patient´s timeline.

Temporal Expressions: Alongside the events, specific temporal expressions are annotated. These could be explicit dates (e.g., "January 15, 2023"), relative times (e.g., "three weeks later"), or even more general temporal indicators (e.g., "during the last visit"). These expressions help to ground the events in time.

Temporal Relations: Temporal annotation focuses on defining relations (TLINKs) between events and temporal expressions, which determine their temporal order.

We restrict our study to a common subset of temporal relation types, as presented in Table 1, in order to precisely capture temporal connections. Specifically, we focus on the following relations: `BEFORE`, where event A occurs prior to event B (e.g., `biopsy-BEFORE-surgery`); `CONTAINS`, where event A fully encompasses event B (`hospital stay-CONTAINS-surgery`); and `OVERLAP`, where event A and event B occur concurrently for some duration (`chemotherapy session-OVERLAP-radiation therapy`, if they were administered at the same time).

4.2 Low-resource Scenarios

Low-resource setting. Our goal is to assess the effectiveness of the data augmentation strategies introduced in Section 3 for clinical temporal relation extraction under conditions of extreme data scarcity. We aim to determine whether these strategies can generate quality training instances that enable models to generalize to complex temporal reasoning tasks in low-resource clinical settings, while minimizing the need for manual annotation. To this end, we design a series of experimental scenarios that simulate varying levels of annotation availability. We consider a zero-resource setting with no annotated documents (0-doc) as a baseline, as well as low-resource conditions with one, two, and up to ten annotated documents (1-doc, 2-doc, and 10-doc). For each data regime, we sample five random subsets to conduct evaluation on golden annotations, ensuring robustness in the evaluation across different document selections. Table 1 shows the average number of gold annotations for each relation label in the different low-resource conditions.

Data-augmentation settings. We generate synthetic data for each low-resource scenario by leveraging the information available in that setting. For each scenario, we apply both **TBG** and **PBG** prompting strategies to construct augmented datasets. All synthetic data are generated using the *Mistral-7B-Instruct-v0.3* model. It is important to note that temporal relations in the clinical domain are highly imbalanced, with a strong predominance of `OUTREL` (i.e., no explicit relation) and `CONTAINS`. To mitigate this issue, for each low-resource scenario, we construct a quasi-balanced synthetic dataset of 25,000 examples; 5,000 instances for each positive label (15,000 in total) and 10,000 instances of `OUTREL`. The negative class (`OUTREL`) is generated by inverting positive relations of type `BEFORE` and `CONTAINS`, yielding 5,000 negative examples each.

0-doc generation To compensate for the absence of an annotated training corpus in 0-doc scenario, we generated a synthetic dataset using SNOMED-CT as a knowledge source. We extracted pairs of clinical events directly from the SNOMED-CT hierarchy. However, because we lacked real-world data to determine their true chronological order, we assigned temporal relations (`OVERLAP`, `BEFORE`, or `CONTAINS`) to these pairs at random. This approach allowed us to fine-tune the model on domain-specific terminology and relationship structures, establishing a baseline performance in a 0-doc setting.

4.3 TRE Approaches

In the classical RE the task is concived as a classification problem. Given a pair of entities $e1$ and $e2$ within the context X, the goal is to maximize the probability assigned to the correct class y (the actual relation). The set of classes includes all predefined relations, along with an additional 'not related' (OUTREL) label, resulting in quadratic growth in the number of possible pairwise hypotheses. We compare two discriminative architectures: a RoBERTa-based **Entity Marker** approach, which extracts features from specific entity boundaries, and a T5-based **sequence classification** approach, which uses the entire sequence representation.

Entity Marker Approach Following (Soares et al., 2019), we implemented an entity marker (EM) approach using an encoder LM. We explicitly insert special tokens (`E0S`, `E0E`, `E1S`, `E1E`) around the target entities to mark their boundaries within the sentence. To capture the contextual representation, we extract the hidden states corresponding to the start markers (`E0S` and `E1S`) from the model´s final layer. These two vectors are concatenated to form a joint representation $h_{pair} = [h_{E0S}; h_{E1S}]$, which is then passed through a dense layer with a *tanh* activation function before the final classification layer. The `xlm-roberta-large` (Zhou and Chen, 2022) model was fine-tuned for 10 epochs with a learning rate of `2e-5` using the AdamW optimizer. We employed a per-device batch size of 16 with gradient accumulation of 4 steps. Further settings included a weight decay of `1e-5`, a constant learning rate scheduler, FP16 mixed precision, and a random seed of 7.

Sequence Classification Approach In addition to the RoBERTa model, we adapted the encoder architecture of T5 (Raffel et al., 2020) (`t5-base`) for sequence classification. Although T5 is originally designed as an encoder-decoder model for text-to-text tasks, we utilize only the **encoder component** in a purely discriminative setup. The model processes the input formatted as "`task: relext; text:` {`text`}", where `text` contains marked entities. The classification head maps the sequence´s contextualized representation to class logits, which are then optimized using a weighted multi-label cross-entropy loss to mitigate class imbalance in the temporal dataset. This approach allows us to compare the effectiveness of T5´s pre-trained encoder weights against the RoBERTa-large baseline in a low-resource clinical setting. Given the class imbalance inherent in temporal datasets, the implementation used the Binary Cross-Entropy as the loss function. The T5 model was trained for 10 epochs with a learning rate of `1e-4` using the Adafactor optimizer. The training configuration used a batch size of 32 with 8 gradient accumulation steps and a constant learning rate scheduler.

4.4 Evaluation

The evaluation framework *macro-averaged* F1 after temporal closure is shared for both approaches, ensuring a fair comparison of their predictive capabilities and their synergy with symbolic reasoning.

Metric Given the high class imbalance, specifically the prevalence of the `OUTREL` category, we primarily report *macro-averaged* F1 scores to ensure that minority temporal relations (i.e., `OVERLAP`) are adequately represented.

Temporal Closure A significant challenge in evaluating temporal RE is ensuring consistency. To address this, we implement a post-resolution step based on Allen's Interval Algebra (Allen, 1983). We map the predicted labels to Allen's 13 basic primitives. The process follows an iterative constraint satisfaction approach:

1. **Symmetry Resolution:** We first resolve inverse relations (e.g., if the model predicts $A \xrightarrow{\texttt{BEFORE}} B$, we automatically infer $B \xrightarrow{\texttt{PRECEDED\_BY}} A$).
2. **Transitive Composition:** We apply the composition operator ($\circ$) across chains of relations. For every triplet entities (A, B, C), if the relations (A, B) and (B, C) are known, we derive the relation (A, C) using a composition table. For example, $A \xrightarrow{\texttt{BEFORE}} B \circ B \xrightarrow{\texttt{BEFORE}} C \implies A \xrightarrow{\texttt{BEFORE}} C$.
3. **Handling `OUTREL` and Ambiguity:** We treat the `OUTREL` label as a "vague" relation (the union of all possible Allen primitives). The solver attempts to reduce this vagueness. If a logical chain implies a specific relation (e.g., `BEFORE`) where the model predicted `OUTREL`, the relation is updated to enhance recall.
4. **Conflict Resolution:** In cases of logical inconsistency (e.g., the model predicts $A \xrightarrow{\texttt{OVERLAP}} C$ but the transitive chain implies $A \xrightarrow{\text{BEFORE}} C$), the engine flags the relation as inconsistent and reverts it to `OUTREL` to maintain document integrity.

This iterative process continues until the temporal graph converges and no further relations can be inferred. This symbolic layer allows the system to correct local model errors and recover relations that were missed during the initial pass, effectively moving

	E3C dataset			THYME dataset		
Encoder	**1-doc**	**2-doc**	**10-doc**	**1-doc**	**2-doc**	**10-doc**
$\text{XLMR}_{\text{GOLD}}$	$2.8_{\pm 1.3}$	$2.9_{\pm 3.2}$	$1.9_{\pm 2.9}$	$3.0_{\pm 3.1}$	$4.1_{\pm 3.7}$	$6.4_{\pm 6.7}$
XLMR_{TBG}	$\mathbf{12.7}_{\pm 4.3}$	$14.3_{\pm 1.3}$	$17.7_{\pm 1.2}$	$\mathbf{23.4}_{\pm 3.0}$	$\mathbf{24.7}_{\pm 1.7}$	$26.7_{\pm 0.6}$
XLMR_{PBG}	$12.3_{\pm 3.5}$	$\mathbf{14.5}_{\pm 1.7}$	$\mathbf{18.4}_{\pm 3.0}$	$21.8_{\pm 4.8}$	$22.7_{\pm 2.3}$	$\mathbf{29.7}_{\pm 1.6}$
T5_{GOLD}	$7.7_{\pm 3.1}$	$9.1_{\pm 1.6}$	$13.4_{\pm 2.6}$	$10.5_{\pm 2.3}$	$11.1_{\pm 1.4}$	$20.1_{\pm 5.7}$
T5_{TBG}	$\mathbf{14.2}_{\pm 3.3}$	$14.0_{\pm 3.2}$	$16.4_{\pm 0.7}$	$\mathbf{21.9}_{\pm 3.9}$	$\mathbf{22.2}_{\pm 2.1}$	$24.4_{\pm 1.4}$
T5_{PBG}	$12.8_{\pm 2.8}$	$\mathbf{14.7}_{\pm 3.4}$	$\mathbf{16.9}_{\pm 1.4}$	$21.9_{\pm 2.0}$	$22.1_{\pm 3.1}$	$\mathbf{28.0}_{\pm 1.4}$

Table 2: In-domain Macro F_1 (Mean ± Std. Dev.) results across varying amounts of annotated documents of the three strategies. GOLD refers to manually annotated documents, whereas TBG and PBG refer to triplet-based generation and paraphrase-based generation, respectively.

from independent pairwise predictions to a coherent global temporal timeline.

5 Results

This section evaluates each DA technique across different data regimes, with results reported for in-domain (Section 5.1) and cross-domain (Section 5.2) settings.

5.1 In-domain Experiments

We first analyze the impact of annotation on synthetic data generation by comparing models trained with augmented data to those relying solely on manually annotated instances. As shown in Table 2, incorporating synthetic data during training consistently improves performance over using only gold annotations (GOLD), regardless of the augmentation strategy (TBG or PBG). These results indicate that annotating even a single document (1-doc) provides sufficient signal to initiate the generation of useful synthetic examples.

In contrast, models trained exclusively on gold data continue to struggle even when up to 10 documents are annotated (corresponding to approximately 300 and 700 positive instances in E3C and THYME, respectively) and exhibit only marginal gains as the amount of annotated data increases.

Table 3 reports the performance of TBG and PBG with 10 annotated documents, comparing them against zero-shot generation (0-doc) and models trained on the full available training set. The results demonstrate the effectiveness of the proposed approaches, highlighting that minimally informed generation is crucial for achieving competitive and well-controlled performance. Furthermore, experiments on E3C underscore the inherent difficulty of the task, as models trained on the full dataset yield only modest improvements over those trained with TBG and PBG.

	E3C		THYME	
	XLMR	**T5**	**XLMR**	$\mathbf{T5_{enc}}$
0-doc	12.7	11.9	14.0	15.1
TBG	17.7	16.4	26.7	24.4
PBG	**18.4**	16.9	29.7	28.0
FT	12.4	**18.9**	**54.1**	**47.1**

Table 3: Macro F_1 results of the baseline (0-doc), TBG and PBG generations with 10 annotated documents, and with full train (FT) for fine-tuning both models in each dataset.

Comparing generation approaches, in the lowest resource setting (1-doc), TBG generation seems more effective than PBG. As discussed in Section 3.1, its weaker contextual constraints enable more diverse and creative outputs (see Section 6). In contrast, paraphrasing produces samples that remain closer to the original instances. However, as more annotated data becomes available (e.g., 10-doc), PBG proves more effective for expanding the training set.

Figure 4 presents per-relation results for XLMR on THYME, highlighting the impact of different augmentation strategies. In low-resource settings, training on gold data alone ($\text{XLMR}_{\text{GOLD}}$) yields poor performance, with a strong bias toward the majority class (`OUTREL`) and failure to capture positive relations (`BEFORE`, `CONTAINS`, `OVERLAP`). In contrast, PBG and TBG reduce this bias and improve performance across relation types, likely due to a more balanced distribution of generated examples. While both approaches perform similarly overall, TBG has a slight advantage in very low-resource settings (1–2

Encoder	THYME → E3C			E3C → THYME		
	1-doc	2-doc	10-doc	1-doc	2-doc	10-doc
$XLMR_{TGB}$	**12.3**$_{\pm 2.6}$	**13.0**$_{\pm 2.7}$	13.0$_{\pm 1.3}$	**20.8**$_{\pm 2.7}$	**21.2**$_{\pm 2.0}$	24.4$_{\pm 0.9}$
$XLMR_{PBG}$	9.9$_{\pm 3.4}$	11.8$_{\pm 3.0}$	**15.9**$_{\pm 0.4}$	17.1$_{\pm 4.3}$	21.0$_{\pm 2.5}$	**26.9**$_{\pm 2.1}$
$T5_{TBG}$	12.1$_{\pm 1.7}$	**13.1**$_{\pm 2.2}$	12.6$_{\pm 0.6}$	**15.0**$_{\pm 2.1}$	16.9$_{\pm 1.2}$	17.5$_{\pm 1.2}$
$T5_{PBG}$	**12.3**$_{\pm 2.4}$	11.1$_{\pm 2.0}$	**14.4**$_{\pm 1.3}$	13.5$_{\pm 2.3}$	**17.3**$_{\pm 1.2}$	**18.6**$_{\pm 1.0}$

Table 4: Cross-domain Macro F1 (Mean ± Std. Dev.) results across varying amounts of annotated documents of the two generation strategies: triplet-based generation TBG and paraphrase-based generation (PBG).

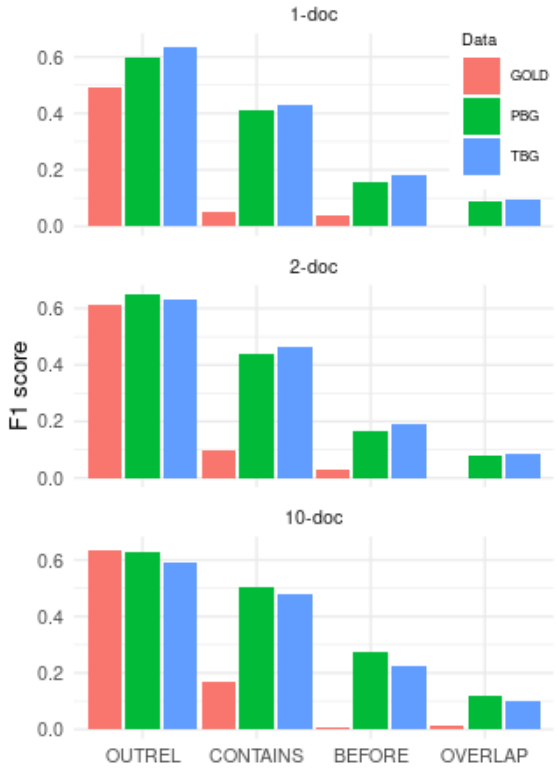


Figure 4: TGB, PBG and GOLD Label-wise comparison on different document amounts.

docs), whereas PBG becomes more effective with more annotated data (e.g., 10 docs), consistent with Table 2.

5.2 Cross-domain Experiments

We analyze how well the generation approaches generalize across datasets under different supervision regimes. Similar to previous experiments, we informed the generation annotating and training with 1, 2, and 10 documents from one dataset (e.g., THYME) and evaluate on the other dataset (e.g., E3C). Table 4 shows the results of the experiments.

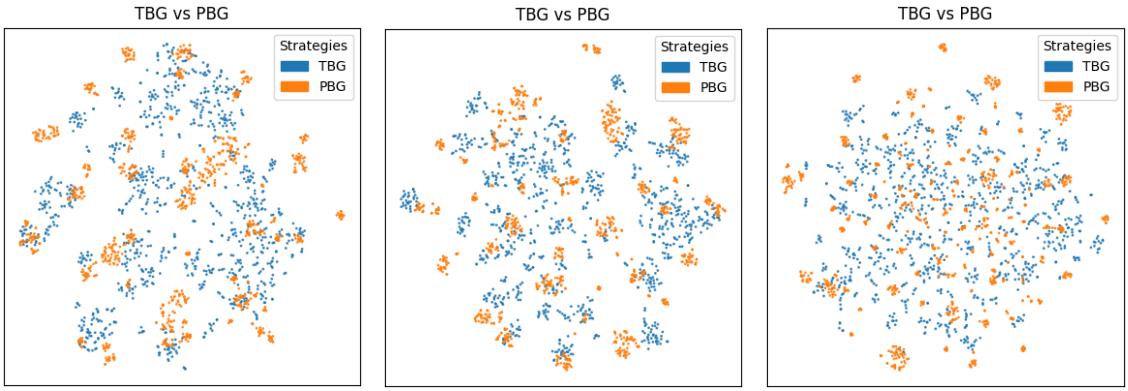
The results in this setting are broadly comparable to those obtained in the in-domain experiments (Table 2), suggesting that generation-based approaches can produce diverse examples that approximate the target distribution and generalize across datasets. As expected, in very low-resource scenarios (1-doc), performance remains similar in both settings. However, as the number of annotated documents increases, a gap emerges between in-domain and cross-domain F-scores. For example, when 10 THYME documents are used for training data generation, $T5_{PBG}$ drops by approximately 10 F-score points (from 28.0 to 18.6) when evaluated on E3C.

In contrast, informed generation exhibits stronger generalization than zero-shot data generation. The 0-doc setting achieves F-scores between 12.8 and 13.9 on E3C and THYME, whereas our approach reaches up to 26.9 F1 when transferring from E3C to THYME. Moreover, generation-based methods achieve transfer performance close to models trained on gold annotations. For instance, PBG with 10 documents attains 15.9 F1 on E3C, compared to 22.3 F1 when XLMR is trained on the full THYME training set; a similar trend is observed in the reverse transfer (26.9 vs. 32.7). Consistent with previous findings, PBG is the most effective strategy as more annotated data becomes available.

6 Discussion

We analyze the impact of TBG and PBG on the diversity/fidelity trade-off (creativity vs semantic preservation) in THYME. Using BLEU, BERTScore, and t-SNE analysis across different annotation regimes (1-, 2-, and 10-doc), we measure lexical and semantic distances between generated and original examples, as well as changes in diversity.

Table 5 summarizes the diversity of the



(a) 1-doc (b) 2-doc (c) 10-doc

Figure 5: t-SNE analysis of random 1000 synthetic examples for 1-doc, 2-doc, and 10-doc.

		BLEU Score	BERT Score	#Uniq. words	#Total words
1-d	PBG	$41.2_{\pm 3.2}$	$86.7_{\pm 1.0}$	2747	523935
	TBG	$6.7_{\pm 2.4}$	$64.7_{\pm 0.6}$	5900	441542
2-d	PBG	$38.7_{\pm 3.5}$	$86.2_{\pm 1.2}$	4669	574899
	TBG	$6.3_{\pm 2.8}$	$65.3_{\pm 0.8}$	6863	427223
10-d	PBG	$39.6_{\pm 5.1}$	$86.2_{\pm 1.8}$	3035	574455
	TBG	$6.2_{\pm 3.3}$	$64.9_{\pm 1.2}$	6219	491624

Table 5: BLEU and BERT Scores (Mean ± Std. Dev.) for 1, 2, 10-doc THYME settings.

generated data in terms of form and meaning. As expected, TBG consistently yields lower BLEU and BERTScore values, indicating greater lexical and semantic variation compared to PBG. These scores remain stable across settings, as each generated instance is evaluated against its corresponding original and results are averaged, making the metrics insensitive to the total number of examples. The higher number of unique words further supports TBG's greater creativity.

Figure 5 illustrates how diversity evolves with increasing amounts of annotated data. Consistent with the BERTScore results, TBG-generated instances (orange colored) form less dense clusters around the original examples than those produced by PBG (blue colored), indicating greater semantic diversity. However, this effect diminishes as the number of annotated instances increases, reflecting the growing intrinsic variability of the original data.

7 *Conclusions and future work*

This work demonstrates that minimal supervision is crucial for achieving both diversity and domain relevance in synthetic CTRE data generation. In contrast, zero-shot (0-doc) approaches often produce content that fails to meet task-specific requirements, resulting in inferior performance. Even a single annotated document substantially improves generation quality by guiding models toward more faithful domain-specific outputs.

In extremely low-resource settings, conditioning generation on entity pairs and relation labels (TBG) obtains a performance similar to that obtained by relying on full contextual input (PBG). The experimental results show that TBG achieves performance comparable to that of PBG in extremely low-resource settings, highlighting its effectiveness under minimal supervision. These findings suggest that it may be possible to successfully adopt an annotation strategy in which clinicians provide prototypical temporal relations for a small set of entity pairs, which can then be expanded through TBG with no initial temporal relation annotation at all. This approach is particularly valuable given the inherent difficulty in involving clinicians in annotation tasks.

In future work intent to investigate the complementarity of these data augmentation strategies and extend the analysis to other large language models.

8 *Limitations*

This study focuses on the three most frequent temporal relations and uses small LLMs due to resource constraints typical in hospital settings. Additionally, dataset licensing restrictions may limit reproducibility and data sharing.

Acknowledgements

This work has been partially supported by European Research Council under Horizon Europe (grant number 10113572, related to LUMINOUS project) the HiTZ Center and the Basque Government, Spain (Research group funding IT1570-22) as well as by MCIN/AEI/10.13039/5011 00011033 Spanish Ministry of Universities, Science and Innovation by means of the projects: EDHIA PID2022-136522OB-C22, HumanAIze AIA2025-163322-C61 and AI4I/MOLVI PID2024-157855OB-C32 (also supported by FEDER, UE).

References

Alfattni, G., N. Peek, and G. Nenadic. 2020. Extraction of temporal relations from clinical free text: A systematic review of current approaches. *Journal of Biomedical Informatics*, page 103488.

Allen, J. F. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Baldini Soares, L., N. FitzGerald, J. Ling, and T. Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2895–2905, Florence, Italy, July. Association for Computational Linguistics.

Chaturvedi, R., P. Baghershahi, S. Medya, and B. Di Eugenio. 2025. Temporal relation extraction in clinical texts: A span-based graph transformer approach. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25765–25788, Vienna, Austria, July. Association for Computational Linguistics.

Cheng, C. and J. C. Weiss. 2023. Typed markers and context for clinical temporal relation extraction. In *Machine Learning for Healthcare Conference*, pages 94–109. PMLR.

Delmas, M., M. Wysocka, and A. Freitas. 2024. Relation extraction in underexplored biomedical domains: A diversity-optimized sampling and synthetic data generation approach. *Computational Linguistics*, 50(3):953–1000, 09.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Dligach, D., T. Miller, C. Lin, S. Bethard, and G. Savova. 2017. Neural temporal relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 746–751.

El Khettari, O., S. Quiniou, and S. Chaffron. 2025. Summarization for generative relation extraction in the microbiome domain. In F. Bechet, A.-G. Chifu, K. Pinel-sauvagnat, B. Favre, E. Maes, and D. Nurbakova, editors, *Actes de l'atelier Traitement du langage médical à l'époque des LLMs 2025 (MLP-LLM)*, pages 68–82, Marseille, France, 6. ATALA \\& ARIA.

Estiri, H., Z. H. Strasser, J. G. Klann, T. H. McCoy, K. B. Wagholikar, S. Vasey, V. M. Castro, M. E. Murphy, and S. N. Murphy. 2020. Transitive sequencing medical records for mining predictive and interpretable temporal representations. *Patterns*, 1(4):100051.

Gao, Y., F. Hou, H. Jahnke, and R. Wang. 2023. Data augmentation with diversified rephrasing for low-resource neural machine translation. In M. Utiyama and R. Wang, editors, *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 35–47, Macau SAR, China, September. Asia-Pacific Association for Machine Translation.

Ghosh, S., C. K. R. Evuru, S. Kumar, S. Ramaneswaran, S. Sakshi, U. Tyagi, and D. Manocha. 2023. DALE: Generative data augmentation for low-resource legal NLP. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8511–8565, Singapore, December. Association for Computational Linguistics.

Jimenez Gutierrez, B., N. McNeal, C. Washington, Y. Chen, L. Li, H. Sun, and Y. Su. 2022. Thinking about GPT-3 in-context learning for biomedical IE? think again. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4497–4512, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Josifoski, M., M. Sakota, M. Peyrard, and R. West. 2023. Exploiting asymmetry for synthetic training data generation: Synthie and the case of information extraction. *arXiv preprint arXiv:2303.04132*.

Knez, T. and S. Žitnik. 2024. Multimodal learning for temporal relation extraction in clinical texts. *Journal of the American Medical Informatics Association*, 31(6):1380–1387, 03.

Labrak, Y., M. Rouvier, and R. Dufour. 2024. A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2049–2066, Torino, Italia, May. ELRA and ICCL.

Lin, C., T. Miller, D. Dligach, S. Bethard, and G. Savova. 2019. A BERT-based universal model for both within- and cross-sentence clinical temporal relation extraction. In A. Rumshisky, K. Roberts, S. Bethard, and T. Naumann, editors, *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71, Minneapolis, Minnesota, USA, June. Association for Computational Linguistics.

Lin, C., T. Miller, D. Dligach, S. Bethard, and G. Savova. 2021. EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain. In D. Demner-Fushman, K. B. Cohen, S. Ananiadou, and J. Tsujii, editors, *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 191–201, Online, June. Association for Computational Linguistics.

Lin, Y., K. Lu, S. Yu, T. Cai, and M. Zitnik. 2023. Multimodal learning on graphs for disease relation extraction. *Journal of Biomedical Informatics*, 143:104415.

Magnini, B., B. Altuna, A. Lavelli, M. Speranza, and R. Zanoli. 2021. The e3c project: European clinical case corpus. *Language*, 1(L2):L3.

Ning, Q., Z. Feng, and D. Roth. 2017. A structured learning approach to temporal relation extraction. In M. Palmer, R. Hwa, and S. Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1027–1037, Copenhagen, Denmark, September. Association for Computational Linguistics.

Ning, Q., S. Subramanian, and D. Roth. 2019. An improved neural baseline for temporal relation extraction. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6203–6209, Hong Kong, China, November. Association for Computational Linguistics.

Olex, A. L. and B. T. McInnes. 2022. Temporal disambiguation of relative temporal expressions in clinical texts. *Frontiers Res. Metrics Anal.*, 7.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Roccabruna, G., M. Rizzoli, and G. Riccardi. 2024. Will LLMs replace the encoder-only models in temporal relation classification? In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20402–20415, Miami, Florida, USA, November. Association for Computational Linguistics.

Ross, J., S. W. Tu, S. Carini, and I. Sim. 2010. Analysis of eligibility criteria com-

plexity in clinical trials. *Summit on Translational Bioinformatics*, 2010:46 – 50.

Rubin, R. B. and M. P. McHugh. 1987. Development of parasocial interaction relationships.

Santamaria, E. A., O. L. de Lacalle, A. Atutxa, and K. Gojenola. 2025. Do entailment models know about reasoning temporal ordering on clinical texts? *Procesamiento del Lenguaje Natural*, 74:349–362.

Sheikhalishahi, S., R. Miotto, J. T. Dudley, A. Lavelli, F. Rinaldi, and V. Osmani. 2019. Natural language processing of clinical notes on chronic diseases: Systematic review. *CoRR*, abs/1908.05780.

Soares, L. B., N. Fitzgerald, J. Ling, and T. Kwiatkowski. 2019. Matching the blanks: Distributional similarity for relation learning. *arXiv preprint arXiv:1906.03158*.

Styler IV, W. F., S. Bethard, S. Finan, M. Palmer, S. Pradhan, P. C. De Groen, B. Erickson, T. Miller, C. Lin, G. Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the association for computational linguistics*, 2:143–154.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, L., Q. Wang, H. Bai, C. Liu, W. Liu, Y. Zhang, L. Jiang, H. Xu, K. Wang, and Y. Zhou. 2020. Ehr2vec: Representation learning of medical concepts from temporal patterns of clinical notes based on self-attention mechanism. *Frontiers in Genetics*, 11.

Wei, J. and K. Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China, November. Association for Computational Linguistics.

Xia, M., X. Kong, A. Anastasopoulos, and G. Neubig. 2019. Generalized data augmentation for low-resource translation. In A. Korhonen, D. Traum, and L. Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796, Florence, Italy, July. Association for Computational Linguistics.

Zhong, Y., Y. Gao, X. Zhang, and H. Wang. 2025. ODDA: An OODA-driven diverse data augmentation framework for low-resource relation extraction. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 267–285, Vienna, Austria, July. Association for Computational Linguistics.

Zhou, W. and M. Chen. 2022. An improved baseline for sentence-level relation extraction. In Y. He, H. Ji, S. Li, Y. Liu, and C.-H. Chang, editors, *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 161–168, Online only, November. Association for Computational Linguistics.

Zhou, Y., Y. Yan, R. Han, J. H. Caufield, K.-W. Chang, Y. Sun, P. Ping, and W. Wang. 2021. Clinical temporal relation extraction with probabilistic soft logic regularization and global inference. In *AAAI*, pages 14647–14655. AAAI Press.

Zhu, X., J. M. Plasek, C. Tang, W. Al-Assad, Z. Zhang, Y. Xiong, L. Wang, S. Yerneni, C. Ortega, M.-J. Kang, et al. 2021. Embedding, aligning and reconstructing clinical notes to explore sepsis. *BMC research notes*, 14(1):1–6.

A Appendix

Relation	TBG	PBG
BEFORE	*head*: entered *tail*: incision `She was E0S entered E0E into the hospital through a surgical E1S incision E1E.`	*head*: intervention *tail*: consent `The decision for the surgical E0S intervention E0E required the patient's E1S consent E1E.`
CONTAINS	*head*: mass *tail*: sent `The biopsy results indicated a malignant E0S mass E0E, confirming that it has been E1S sent E1E for further pathological examination.`	*head*: mass *tail*: extended `Antero-posterior and lateral knee X-rays demonstrated degenerative changes and a large, calcified E0S mass E0E that began below the patella and E1S extended E1E up to the suprapatellar area.`
OVERLAP	*head*: range *tail*: movements `Upon physical examination, the E1S movements E1E demonstrate a limited E0S range E0E of the right upper extremity, with reduced ability to flex and extend the elbow and wrist.`	*head*: complaints *tail*: pain `There was no mention of E1S pain E1E or any reported E0S complaints E0E.`

Table 6: E3C generation examples.

Relation	TBG	PBG
BEFORE	*head*: radiochemotherapy *tail*: tumor `The patient was prescribed to undergo E0S radiochemotherapy E0E prior to any growth or E1S tumor E1E progression, as per the treatment plan.`	*head*: radiochemotherapy *tail*: resolution `On February 22, 2010, a CT-scan of the abdomen and pelvis indicated the previously enlarged perirectal node had shown E1S resolution E1E following new adjuvant E0S radiochemotherapy E0E.`
CONTAINS	*head*: now *tail*: involved `Currently, the patient presents with E0S now E0E demonstrates E1S involved E1E aortic regurgitation as evidenced by the echocardiogram.`	*head*: surgery *tail*: ileostomy `On February 25, 2010, Dr. Dixon performed E0S surgery E0E for a low-anterior resection and created a protective E1S ileostomy E1E for protection.`
OVERLAP	*head*: pain *tail*: complained `The patient E1S complained E1E of severe abdominal E0S pain E0E that radiated to her back, which awaits more diagnostic assessment.`	*head*: lymph nodes *tail*: involvement `Therefore, the conclusion emphasizes that the 17 regional E0S lymph nodes E0E did not show any sign of negative E1S involvement E1E by the tumor.`

Table 7: THYME generation examples.

Label	Verbalization templates	
BEFORE	{e1} occurs before {e2}	{e2} occurs after {e1}
	{e1} happens before {e2}	{e2} happens after {e1}
CONTAINS	{e1} reports {e2}	{e1} is observed in {e2}
	{e1} demonstrates {e2}	{e1} objectifies {e2}
	{e1} shows {e2}	{e1} confirms {e2}
OVERLAP	{e1} and {e2} occur at same time at some point in time	
	{e1} and {e2} overlap at some point in time	

Table 8: TBG Verbalization templates.

	0-doc	1-doc	2-doc	10-doc	FT
In Domain Encoder E3C					
XLMR_G	-	$2.8_{\pm 1.3}$	$2.9_{\pm 3.2}$	$1.9_{\pm 2.9}$	12.4
XLMR_T	12.7	$\mathbf{12.7}_{\pm 4.3}$	$14.3_{\pm 1.3}$	$17.7_{\pm 1.2}$	-
XLMR_P	-	$12.3_{\pm 3.5}$	$\mathbf{14.5}_{\pm 1.7}$	$\mathbf{18.4}_{\pm 3.0}$	-
T5_G	-	$7.7_{\pm 3.1}$	$9.1_{\pm 1.6}$	$13.4_{\pm 2.6}$	18.9
T5_T	11.9	$\mathbf{14.2}_{\pm 3.3}$	$14.0_{\pm 3.2}$	$16.4_{\pm 0.7}$	-
T5_P	-	$12.8_{\pm 2.8}$	$\mathbf{14.7}_{\pm 3.4}$	$\mathbf{16.9}_{\pm 1.4}$	-
In Domain Encoder THYME					
XLMR_G	-	$3.0_{\pm 3.1}$	$4.1_{\pm 3.7}$	$6.4_{\pm 6.7}$	54.1
XLMR_T	14.0	$\mathbf{23.4}_{\pm 3.0}$	$\mathbf{24.7}_{\pm 1.7}$	$26.7_{\pm 0.6}$	-
XLMR_P	-	$21.8_{\pm 4.8}$	$22.7_{\pm 2.3}$	$\mathbf{29.7}_{\pm 1.6}$	-
T5_G	-	$10.5_{\pm 2.3}$	$11.1_{\pm 1.4}$	$20.1_{\pm 5.7}$	47.1
T5_T	15.1	$\mathbf{21.9}_{\pm 3.9}$	$\mathbf{22.2}_{\pm 2.1}$	$24.4_{\pm 1.4}$	-
T5_P	-	$21.9_{\pm 2.0}$	$22.1_{\pm 3.1}$	$\mathbf{28.0}_{\pm 1.4}$	-

Table 9: In-domain Macro F_1 (Mean ± Std. Dev.) results across varying volumes of annotated documents. Results for three strategies: Gold (G), TBG (T), and PBG (P). FT denotes the Full Train regime, while 0-doc stands the Baseline.

	0-doc	1-doc	2-doc	10-doc
Transfer Encoder THYME to E3C				
XLMR_T	12.8	$\mathbf{12.3}_{\pm 2.6}$	$\mathbf{13.0}_{\pm 2.7}$	$13.0_{\pm 1.3}$
XLMR_P	-	$9.9_{\pm 3.4}$	$11.8_{\pm 3.0}$	$\mathbf{15.9}_{\pm 0.4}$
T5_T	13.3	$12.1_{\pm 1.7}$	$\mathbf{13.1}_{\pm 2.2}$	$12.6_{\pm 0.6}$
T5_P	-	$\mathbf{12.3}_{\pm 2.4}$	$11.1_{\pm 2.0}$	$\mathbf{14.4}_{\pm 1.3}$
Transfer Encoder E3C to THYME				
XLMR_T	14.3	$\mathbf{20.8}_{\pm 2.7}$	$\mathbf{21.2}_{\pm 2.0}$	$24.4_{\pm 0.9}$
XLMR_P	-	$17.1_{\pm 4.3}$	$21.0_{\pm 2.5}$	$\mathbf{26.9}_{\pm 2.1}$
T5_T	13.9	$\mathbf{15.0}_{\pm 2.1}$	$16.9_{\pm 1.2}$	$17.5_{\pm 1.2}$
T5_P	-	$13.5_{\pm 2.3}$	$\mathbf{17.3}_{\pm 1.2}$	$\mathbf{18.6}_{\pm 1.0}$

Table 10: Cross-domain transferred Macro F_1 (Mean ± Std. Dev.) results across varying volumes of annotated documents. Results for three strategies: TBG (T) and PBG (P). 0-doc stands the Baseline.